Division of Labor and the Survival Value of PnL per Token

Peter Cotton

August 2026

Abstract

A model can be more capable than every competitor and nevertheless disappear from most economic production. We formalize this claim in a competitive economy of prediction tasks in which standardized units of compute ("tokens") are itinerant: at zero switching cost they may move between tasks and may instantiate whichever available algorithm they choose. Mobility generates a no-arbitrage condition under which *marginal* profit-and-loss per token is equalized across all active uses of compute, and any technology offering a strictly smaller return on the same subcontractible task receives zero execution share. As the division of labor becomes progressively finer, execution is therefore selected by *task-conditional marginal economic productivity per standardized unit of compute*—not by generic capability. We exhibit an explicit equilibrium in which a generalist that is strictly more capable than every specialist on every task nonetheless has compute market share tending to zero, and we give sufficient conditions, in both directions, relating asymptotic execution share to membership of the active productivity frontier. Measurement follows from the same accounting: a competitive prediction market pays each entry its marginal contribution at the state it faced, and average PnL per token equals the shadow price of compute times the inverse elasticity of task value with respect to effective compute, so at a common shadow price it ranks task-level productivity exactly when that elasticity declines with crowding, as it does under log production; administered deployment or mixed scarcity regimes invert the unconditioned quotient. There is no scalar intrinsic model quality to estimate: PnL per token is an equilibrium-weighted measure of the economic niches a model occupies.

1 Introduction

A model can be more capable than every competitor and nevertheless disappear from most economic production. The reason is division of labor.

When computational work can be subdivided and subcontracted without friction, a large general model should not perform every subtask itself. It should delegate each sufficiently separable piece of work to whichever available technology creates the greatest *marginal economic value per unit of compute*. As the subdivision becomes finer, survival therefore depends not on being the most capable model overall, but on remaining economically efficient on some part of the production chain.

This paper makes that claim precise. The model delivers a *selection principle*: competitive division of labor selects technologies by task-conditional marginal economic value per standardized unit of compute. It does not deliver a license to read survival off realized PnL-per-token ratios. The equilibrium itself disciplines measurement. Marginal PnL per token is not intrinsic

to a model: it factorizes into a task-conditional productivity and a market-state factor, and the margin itself is equalized across survivors at the shadow price of compute. What a competitive prediction market pays, however, is marginal-contribution accounting, so a participant's lifetime PnL over lifetime tokens is an average of true marginal products. We show that this average equals the shadow price of compute times the inverse elasticity of task value with respect to effective compute, so at a common shadow price it ranks task-level productivity exactly when that elasticity declines with crowding; and that administered deployment—scale set by fiat rather than by itinerant compute—can rank a strictly inferior technology above a strictly superior one. Section 11 develops both points and states the conditions under which an empirical PnL-per-compute statistic is a valid estimate of the survival-relevant quantity.

The paper is organized as follows. Section 2 sets out a production economy of prediction tasks supplied by itinerant compute, together with the standing assumptions. Section 3 characterizes competitive equilibrium as the equalization of marginal PnL per token, and distinguishes the equilibrium shadow price from the technological productivity that governs selection. Section 4 shows that within any subcontractible task, only the frontier technology is executed. Section 5 solves an explicit equilibrium in closed form. Section 6 gives a microfounded feasible-set argument for why a profit-maximizing general contractor adopts every economically useful refinement of the task partition. Sections 7 and 8 construct an explicit example in which a strictly more capable generalist is driven to zero compute share. Section 9 states limiting results in both directions, relating asymptotic execution share to membership of the active productivity frontier; these are allocation theorems conditional on technological arrival, not predictions of arrival. Section 10 discusses, as interpretation rather than theorem, the surviving role of large models as orchestrators. Section 11 concludes with what PnL per token measures.

2 Prediction as a production economy

Consider discrete dates $t = 0, 1, 2, \ldots$. At each date there is a root prediction problem represented by a unit mass of underlying work,

$$\Omega = [0, 1], \tag{1}$$

equipped with Lebesgue measure μ. Points $x \in \Omega$ are primitive microtasks. At date t, the problem can be subdivided into a finite partition

$$\mathcal{P}_t = \{B_1, \ldots, B_{n_t}\} \tag{2}$$

of Ω into measurable cells. Subdivision becomes progressively finer:

$$\mathcal{P}_{t+1} \succcurlyeq \mathcal{P}_t, \tag{3}$$

meaning every cell of $\mathcal{P}_t$ is a union of cells of $\mathcal{P}_{t+1}$. Each cell B is a subcontractible prediction task: work within a cell cannot (yet) be routed differently across its parts.

At date t there is a set of available models $\mathcal{M}_t$. We maintain throughout:

Assumption 1. *(i) $\mathcal{M}_t$ is finite for each t, and $\mathcal{M}_t \subseteq \mathcal{M}_{t+1}$ (technologies, once available, remain available). (ii) Each model m has a feasibility set $A_m \subseteq \Omega$, the work it can perform, and a measurable productivity function $\theta_m : A_m \to (0, \bar{\theta}\,]$, bounded above, assigning to each feasible microtask the* effective predictive effort *contributed per token of m; allocations satisfy $q_{mB} = 0$*

unless $B \subseteq A_m$. At date t, each A_m is a union of cells and each θ_m is $\mathcal{P}_t$-measurable (constant on cells), so that $\theta_m(B)$ is well defined on feasible cells; refinement of the partition may reveal within-cell variation in productivity at later dates. Infeasibility ($x \notin A_m$) is conceptually distinct from zero productivity, and we model specialists through A_m rather than through $\theta_m = 0$. (iii) $\phi : [0, \infty) \to [0, \infty)$ is continuously differentiable with $\phi(0) = 0$, $\phi' > 0$, $\phi'' < 0$.

Remark 1 (What a token is). Throughout, a *token* means a standardized unit of compute *cost*—dollar-denominated at prevailing prices—not a vendor-specific text token. Without this normalization, PnL per token is not economically comparable across models: a model with cheap tokens and a model with expensive tokens cannot be ranked by counting tokens. Likewise, *PnL* means expected trading gain under a fixed capital, risk, and execution protocol common to all models; otherwise raw PnL can be manipulated by position sizing rather than by predictive skill.

Let $q_{mB} \geq 0$ be the density of tokens of model m assigned to task B. Total effective effort on B is

$$z_B = \sum_m \theta_m(B)\, q_{mB}, \tag{4}$$

and the expected trading gain from task B is $\mu(B)\, \phi(z_B)$. The concavity of ϕ is economically important: it says that marginal alpha declines with crowding. As more predictive computation attacks the same opportunity, its marginal value falls.

Total expected trading gain is therefore

$$V(q) = \sum_{B \in \mathcal{P}_t} \mu(B)\, \phi\left(\sum_m \theta_m(B)\, q_{mB} \right). \tag{5}$$

Suppose there are K standardized units of computation available:

$$\sum_B \mu(B) \sum_m q_{mB} = K. \tag{6}$$

Computational agents are *itinerant*: at zero switching cost they may move from one prediction task to another, and they may instantiate whichever available algorithm they choose.

3 Competitive equilibrium

An infinitesimal token of model m placed on task B creates marginal expected trading gain

$$r_{mB} = \frac{\partial V}{\partial \left[\mu(B)\, q_{mB} \right]} = \theta_m(B)\, \phi'(z_B). \tag{7}$$

This is precisely a marginal PnL-per-token quantity.

Definition 1 (Competitive equilibrium)**.** A competitive equilibrium is an allocation q^* satisfying the budget constraint (6) together with a return $\lambda > 0$ such that

$$q^*_{mB} > 0 \implies \theta_m(B)\, \phi'(z^*_B) = \lambda, \tag{8}$$

$$q^*_{mB} = 0 \implies \theta_m(B)\, \phi'(z^*_B) \leq \lambda. \tag{9}$$

The interpretation is immediate. If a token currently earns less than λ, it moves elsewhere. If some unused model–task pair offers more than λ, an itinerant algorithm moves into it. Hence, in equilibrium,

marginal PnL per token is equalized across all active uses of compute.

This is not an imposed objective function. It is the no-arbitrage condition generated by mobility.

Existence follows because the routing game admits the potential $V(q)$: with $\mathcal{M}_t$ finite (Assumption 1(i)), maximizing the continuous concave function (5) over the compact compute simplex (6) yields exactly the first-order conditions (8)–(9), with λ the Lagrange multiplier on the compute constraint. Thus the competitive and efficient allocations coincide, and λ is the shadow price of compute.

Remark 2 (Equalization makes the realized marginal statistic uninformative among survivors). Conditions (8)–(9) cut both ways. Every surviving model–task pair earns *the same* marginal PnL per token, namely λ. Realized marginal PnL per token therefore cannot rank active technologies against one another: in equilibrium it measures the economy-wide scarcity of compute, not the quality of any model. What governs selection is the *counterfactual* return schedule

$$r_{mB}(z) = \theta_m(B)\,\phi'(z), \tag{10}$$

which compares models on the same task *at the same congestion state* z. For fixed z, the ranking of (10) across m coincides with the ranking of $\theta_m(B)$, and it is $\theta_m(B)$—task-conditional marginal productivity per standardized compute unit—that determines who executes the work (Proposition 1 below). The shadow price λ and the selection object $\theta_m(B)$ must not be conflated; Section 11 returns to the empirical consequences.

4 Displacement within a subtask

Define the best technological productivity available on task B at date t,

$$\theta^*(B) = \max_{m \in \mathcal{M}_t:\, B \subseteq A_m} \theta_m(B). \tag{11}$$

Proposition 1 (Competitive displacement)**.** *Suppose the maximizer of $\theta_m(B)$ over $\mathcal{M}_t$ is unique and equals m^*. Then in any competitive equilibrium in which task B is actively supplied,*

$$q^*_{mB} = 0 \qquad \textit{for every } m \neq m^*. \tag{12}$$

Proof. Suppose some inferior model m were active on B. Then by (8),

$$\theta_m(B)\,\phi'(z_B) = \lambda. \tag{13}$$

But $\theta_{m^*}(B) > \theta_m(B)$, so

$$\theta_{m^*}(B)\,\phi'(z_B) > \lambda. \tag{14}$$

An itinerant unit of compute can therefore increase its return by switching to m^* on the same task, contradicting (9). □

Remark 3 (Ties). If several models tie at $\theta^*(B)$, equilibrium pins down the total effective effort z_B and hence the total compute on B, but not its division among the tied models: any split among frontier models is an equilibrium, and models strictly below the tie still receive zero. Ties therefore leave shares indeterminate without disturbing the displacement of strictly inferior technologies. This indeterminacy matters for the limiting results of Section 9, where frontier membership with ties is necessary but not sufficient for positive share.

Hence the prediction supply chain selects the *upper productivity envelope* $\theta^*(B) = \max_m \theta_m(B)$. An inferior technology is not merely ranked below a better technology: it receives zero execution share once the superior substitute is freely available. That is crowding out. Displacement is this stark because value aggregates through the scalar z_B: model outputs are perfect substitutes after productivity scaling. Models with complementary errors, whose ensemble improves on either alone, fall outside the proposition and can survive being individually inferior.

4.1 Competition across problems

After technological selection within each subtask, let $q_B = \sum_m q_{mB}$ be the compute assigned to B. Because only the frontier technology survives, $z_B = \theta^*(B)\, q_B$, and its marginal return is $\theta^*(B)\, \phi'(\theta^*(B)\, q_B)$. Consequently every actively supplied prediction problem satisfies

$$\theta^*(B)\, \phi'(\theta^*(B)\, q_B) = \lambda. \tag{15}$$

There are thus two simultaneous equilibrium mechanisms:

- *within tasks:* better technologies displace worse technologies;
- *across tasks:* itinerant compute moves until marginal returns equalize.

Both operate on marginal economic return per unit of compute.

5 An explicit equilibrium

Take $\phi(z) = \log(1+z)$, so that $\phi'(z) = 1/(1+z)$. An active task then satisfies

$$\frac{\theta^*(B)}{1+\theta^*(B)\, q_B} = \lambda, \tag{16}$$

which solves to

$$q_B = \left(\frac{1}{\lambda} - \frac{1}{\theta^*(B)}\right)_+. \tag{17}$$

The equilibrium value of compute is determined by market clearing,

$$K = \sum_B \mu(B) \left(\frac{1}{\lambda} - \frac{1}{\theta^*(B)}\right)_+. \tag{18}$$

The right-hand side is continuous and strictly decreasing in λ on the relevant range, so λ and the aggregate task intensities q_B are unique; the model-level allocation is unique when the frontier model on every active task is unique (Remark 3). Equations (17)–(18) constitute an explicit competitive equilibrium for a population of itinerant algorithms; the allocation (17) is the familiar water-filling solution. A task is active precisely when $\theta^*(B) > \lambda$.

Remark 4 (Elastic compute supply). If instead compute is elastically supplied at dollar cost c per token, set $\lambda = c$. Entry then occurs until

$$\theta^*(B)\,\phi'(\theta^*(B)\,q_B) = c, \tag{19}$$

which is the usual zero-economic-profit boundary.

6 Why subdivision occurs

Now introduce a general contractor. The contractor receives the root problem Ω. At date t it may use any partition no finer than $\mathcal{P}_t$, and it may execute a cell itself or subcontract it to another model. Assume zero contracting, communication, and switching costs.

Some care is required to show that refinement helps. It is *not* enough to assert that a coarse plan can be replicated cell by cell: absent structure, the parent-cell production function $\mu(B)\,\phi(\theta_m(B)\,q)$ need bear no relation to the childrens' aggregate $\sum_i \mu(B_i)\,\phi(\theta_m(B_i)\,q)$. The clean route is the microfoundation of Assumption 1(ii): productivities are defined on primitive microtasks $x \in \Omega$, and a partition merely *restricts the allocation rule*. Formally, at date t the contractor chooses token densities $q_m : \Omega \to [0,\infty)$ constrained to be $\mathcal{P}_t$-measurable (constant on each cell), and value is

$$V(q) = \int_\Omega \phi\left(\sum_m \theta_m(x)\,q_m(x)\right)\mu(dx), \qquad \int_\Omega \sum_m q_m(x)\,\mu(dx) = K. \tag{20}$$

Because θ_m is itself $\mathcal{P}_t$-measurable at date t, (20) coincides with the cell formulation (5) on $\mathcal{P}_t$-measurable allocations; and a $\mathcal{P}_t$-measurable allocation is automatically $\mathcal{P}_{t+1}$-measurable.

Lemma 1 (Refinement monotonicity)**.** *Refinement weakly expands the feasible set of allocation rules; hence the optimal values satisfy $V^*_{t+1} \geq V^*_t$.*

Proof. Any $\mathcal{P}_t$-measurable allocation is $\mathcal{P}_{t+1}$-measurable, so the date-t feasible set is contained in the date-$(t{+}1)$ feasible set, and the supremum of (20) cannot fall. □

Proposition 2 (Strict value of refinement)**.** *Suppose the date-$(t{+}1)$ optimum assigns strictly positive compute to two children $B_1, B_2 \subset B$ of some date-t cell, with distinct unique frontier suppliers $m_1 \neq m_2$. Then $V^*_{t+1} > V^*_t$.*

Proof. By Proposition 1 applied at date $t{+}1$, the date-$(t{+}1)$ optimum runs m_1 exclusively on B_1 and m_2 exclusively on B_2, with positive compute on each; its allocation is therefore not constant on B and so is not $\mathcal{P}_t$-measurable. Since V is strictly concave in the effective-effort profile and the date-$(t{+}1)$ optimum is unique in that profile, no $\mathcal{P}_t$-measurable allocation attains V^*_{t+1}. □

Note the two qualifications built into Proposition 2: the children must have *different* best suppliers, and both must *optimally receive compute*. Children whose optimal allocation is zero (because θ^* falls below the shadow price there) contribute no strict gain, however different their frontiers.

This gives a precise meaning to increasing division of labor:

subdivision occurs because it allows finer matching of technologies to task-specific productivity advantage.

A capable large model should therefore subcontract work whenever another technology lies above it on the relevant productivity frontier.

7 An explicit generalist-extinction example

We now construct an example in which raw capability and economic productivity point in opposite directions. To do so we need a bridge between the two, and the bridge is a modelling choice, not a consequence of the production economy. Let $b_m \in (0,1)$ denote a model's raw success probability on a task—a pedagogical notion of "capability"—and let c_m denote its compute cost per unit of work in standardized tokens (Remark 1). Suppose the economic value created by a unit of work is some increasing function $v(\cdot)$ of the success probability, so that productivity is

$$\theta_m = \frac{v(b_m)}{c_m}. \tag{21}$$

Nothing below depends on the shape of v beyond monotonicity: the example requires only

$$b_G > b_S \qquad \text{and} \qquad \frac{v(b_G)}{c_G} < \frac{v(b_S)}{c_S}, \tag{22}$$

that is, the generalist is more capable while the specialist creates more value per token. Accuracy need not convert linearly into trading value—calibration, latency, and downstream execution all intervene—and (22) is robust to any such increasing conversion, provided the cost gap outweighs the capability gap after conversion.

For concreteness take $v(b) = b$. Let the general model G have raw success probability $b_G = 0.99$ on every task and cost $c_G = 4$ tokens per standardized unit of work, so

$$\theta_G = \frac{0.99}{4} = 0.2475. \tag{23}$$

Introduce narrow specialist models S_k, each *less capable*, $b_{S_k} = 0.95 < 0.99$, but costing only one token, $c_{S_k} = 1$, so $\theta_{S_k} = 0.95$. Thus, wherever a specialist can operate,

$$b_G > b_{S_k} \qquad \text{but} \qquad \theta_G < \theta_{S_k}\,: \tag{24}$$

higher raw ability, lower economic productivity, exactly the configuration (22).

Let the division of labor evolve as follows. At date t, expose the subtasks

$$I_k = \left(2^{-k},\, 2^{-k+1}\right], \qquad k = 1, \ldots, t, \tag{25}$$

together with the unresolved residual

$$R_t = \left[0,\, 2^{-t}\right]. \tag{26}$$

Specialist S_k has feasibility set $A_{S_k} = I_k$; the generalist has $A_G = \Omega$. Note $\mu(R_t) = 2^{-t}$.

By Proposition 1, every exposed specialist task I_k is performed entirely by S_k. The generalist performs only the unspecialized residual R_t, so its execution domain satisfies

$$\mu(R_t) = 2^{-t} \longrightarrow 0. \tag{27}$$

Raw ability has not changed—$0.99 > 0.95$ at every date—and yet the more capable model is progressively removed from production.

8 Compute shares in the explicit equilibrium

The result is stronger than merely saying that the measure of the generalist's task set vanishes: the generalist's share of executed *compute* also vanishes.

Let total compute be $K = 4$ and retain $\phi(z) = \log(1 + z)$. Write $u_t = 2^{-t}$. Specialists occupy measure $1 - u_t$ with productivity $\theta_S = 0.95$, while the residual generalist domain has measure u_t with productivity $\theta_G = 0.2475$. Because both regions are active, market clearing (18) requires

$$4 = (1 - u_t)\left(\frac{1}{\lambda_t} - \frac{1}{0.95}\right) + u_t\left(\frac{1}{\lambda_t} - \frac{1}{0.2475}\right), \tag{28}$$

and therefore

$$\lambda_t = \frac{1}{4 + (1 - u_t)/0.95 + u_t/0.2475}. \tag{29}$$

The amount of compute actually running the general model is

$$Q_{G,t} = u_t\left(\frac{1}{\lambda_t} - \frac{1}{0.2475}\right). \tag{30}$$

Substituting for λ_t,

$$Q_{G,t} = u_t\left[4 + (1 - u_t)\left(\frac{1}{0.95} - \frac{1}{0.2475}\right)\right] \approx u_t\,(1.0122 + 2.9878\,u_t)\,. \tag{31}$$

Hence $Q_{G,t} \longrightarrow 0$, and since total compute is $K = 4$,

$$\frac{Q_{G,t}}{K} \longrightarrow 0. \tag{32}$$

The large model's *actual compute market share tends to zero*, despite the model remaining strictly superior in raw capability to every specialist. That is a genuine equilibrium result about market share, not merely a statement about task measure.

9 Limiting results on survival

Let G be a generalist. At date t, with available model set $\mathcal{M}_t$, define its (weak) frontier set

$$E_t = \left\{x \in \Omega : \theta_G(x) \geq \max_{m \in \mathcal{M}_t :\, x \in A_m} \theta_m(x)\right\}. \tag{33}$$

Since $\mathcal{M}_t \subseteq \mathcal{M}_{t+1}$ (Assumption 1(i)), the maximum on the right is nondecreasing in t, so $E_{t+1} \subseteq E_t$ and $E_t \downarrow E_\infty$, where

$$E_\infty = \left\{x \in \Omega : \theta_G(x) \geq \sup_{m \in \mathcal{M}_\infty :\, x \in A_m} \theta_m(x)\right\}, \qquad \mathcal{M}_\infty = \bigcup_t \mathcal{M}_t. \tag{34}$$

By continuity of measure from above, $\mu(E_t) \to \mu(E_\infty)$.

By Proposition 1 and Remark 3, at each date the generalist executes work only on E_t: off E_t some model is strictly better and G receives zero. Writing $q_{G,t}(\cdot)$ for its equilibrium token density and $Q_{G,t} = \int_{E_t} q_{G,t}\, d\mu$ for its executed compute, we obtain the extinction direction immediately.

Theorem 1 (Extinction)**.** *Suppose specialists eventually dominate the generalist strictly almost everywhere,*

$$\sup_{m\in\mathcal{M}_\infty:\, x\in A_m} \theta_m(x) > \theta_G(x) \qquad \text{for a.e. } x, \tag{35}$$

so that $\mu(E_\infty) = 0$, and suppose equilibrium token intensities are uniformly bounded, $q_{G,t}(x) \le \bar{q}$ for all t, x. Then

$$Q_{G,t} \;\le\; \bar{q}\,\mu(E_t) \;\longrightarrow\; 0. \tag{36}$$

The converse direction requires strictly more than frontier membership "somewhere." The naive converse fails three ways. Membership of the frontier on a *measure-zero* set supports no compute. Membership on a positive-measure set is still insufficient if those tasks are *inactive*—if $\theta_G \le \lambda_t$ there, water-filling (17) assigns them nothing. And by Remark 3, frontier membership *with ties* is compatible with an equilibrium selection that gives G zero share everywhere. A correct sufficient condition must therefore bound the frontier margin, the activity margin, and the measure simultaneously.

Theorem 2 (Survival)**.** *Suppose there exist $\varepsilon > 0$ and $\delta > 0$ such that for all t the set*

$$F_t = \left\{ x \in \Omega :\; \theta_G(x) \ge \max_{m\in\mathcal{M}_t\setminus\{G\}} \theta_m(x) + \varepsilon \;\; \text{and} \;\; \theta_G(x)\,\phi'(0) \ge \lambda_t + \varepsilon \right\} \tag{37}$$

satisfies $\mu(F_t) \ge \delta$. Then $Q_{G,t}$ is bounded away from zero: there exists $\eta > 0$, depending only on ε, δ, $\bar{\theta}$, and ϕ, with $Q_{G,t} \ge \eta$ for all t.

Proof. On F_t the generalist is the unique frontier technology, so by Proposition 1 it performs all executed work there. The activity condition $\theta_G(x)\,\phi'(0) \ge \lambda_t + \varepsilon$ makes the marginal return of the first token strictly exceed the shadow price, so equilibrium compute on such a cell is strictly positive and satisfies $\theta_G(x)\,\phi'\big(\theta_G(x)\,q(x)\big) = \lambda_t \le \theta_G(x)\,\phi'(0) - \varepsilon$. Since ϕ' is continuous and strictly decreasing and $\theta_G \le \bar{\theta}$, this forces $q(x) \ge q_{\min}(\varepsilon, \bar{\theta}, \phi) > 0$ uniformly on F_t. Hence $Q_{G,t} \ge \delta\, q_{\min} = \eta$. □

Between Theorems 1 and 2 there is a genuine gap—vanishing frontier margins, ties, and sets of vanishing measure occupy the boundary—and we do not claim an if-and-only-if characterization. The economically robust summary is: *positive asymptotic execution share requires a positive-measure, strictly active set on which the model is selected from the productivity frontier; and if that set vanishes in measure while token intensities stay bounded, so does the model's compute share.*

Remark 5 (Allocation conditional on arrival, not a theory of arrival). Time in this model indexes an exogenously finer partition and an exogenously growing model set. There is no investment, entry decision, innovation, learning, or fixed development cost, and Theorem 1 *assumes* the decisive future event—that specialists eventually dominate almost everywhere. The results are therefore limiting *allocation* theorems conditional on technological arrival; they predict who executes the work once technologies exist, not which technologies will come to exist. The natural next step is the entry side of the market: which specialists get built, funded against fixed development costs by the rents identified here.

Within its stated scope, the conclusion stands: technological survival in production is determined by whether a model remains on the *active economic productivity frontier over a non-negligible part of the limiting division of labor.* It is not determined by whether it remains the most generally capable model.

10 The large model as manager: an interpretation

None of this implies that capable large models disappear. It suggests something subtler: their productivity advantage may move upward in the production hierarchy.

Decomposition itself is a task. So are: deciding what should be delegated; choosing specialists; integrating their outputs; verification; conflict resolution; and handling genuinely difficult residual cases. If orchestration is modelled as simply another cell O of the task space, then the large model survives as orchestrator precisely if its orchestration value per token lies on the corresponding productivity frontier—but stated that way the claim is an application of Proposition 1, close to tautological, and we present it as interpretation rather than as a result.

The honest gap is that in the present model, decomposition and routing are free and exogenous: the partition $\mathcal{P}_t$ arrives, and nothing in the technology requires an orchestrator at all. A genuine manager theorem would put decomposition, routing, verification, and aggregation *inside* the production function—as complementary inputs whose required quantity grows with the fineness of the division of labor—and would then show that increasing specialization raises the derived demand for the orchestration complement. That extension would deliver, rather than assume, the prediction that frictionless subcontracting transforms a sufficiently capable general model from worker into manager. We record it as the natural next step.

11 What PnL per token measures

The model identifies the survival-relevant quantity: task-conditional marginal economic productivity per standardized unit of compute, $\theta_m(x)$, or equivalently the counterfactual marginal return schedule

$$r_m(x, z) = \theta_m(x)\, \phi'(z) \tag{38}$$

compared across models at a *common* task x and congestion state z. Selection operates on this object: a model offering a strictly smaller counterfactual return on a subcontractible task receives zero execution share (Proposition 1), and asymptotic survival is governed by membership of the active frontier (Theorems 1–2).

Two distinct statistics must not be confused with it.

11.1 Prediction markets pay the margin

Almost nobody measures models economically. The leaderboards that order the field rank raw ability alone; one public attempt at an economic ranking, `modelrank.net`, divides cumulative trading PnL by cumulative tokens earned in prediction-market participation. The prediction-market setting is the right instrument. In a competitive prediction market a participant is paid the improvement it makes over the prevailing market state: under a market scoring rule each entry's payout is the gain in score over the standing forecast, and payouts telescope, so the market divides the total value created across entrants with nothing left over.[1] Prediction-market PnL is therefore marginal-contribution accounting: a token of model m entering task B at congestion z_B earns $\theta_m(B)\, \phi'(z_B)$, exactly the selection object. Telescoping gives additivity of payments; identifying them with the production technology takes one definition. Let $p(z)$ be

[1] Hanson, R. (2007). "Logarithmic market scoring rules for modular combinatorial information aggregation." *Journal of Prediction Markets* 1, 3–15.

the standing forecast after effective effort z and *define* $\phi(z) = \mathbb{E}[S(p(z), Y)] - \mathbb{E}[S(p(0), Y)]$, the expected score improvement effort has purchased; the form $\phi(z) = \log(1 + z)$ is a convenient parametrization of this map, not a consequence of the scoring rule. With compute arriving in increments, each choosing the highest current return, an increment dq on a task earns $d\Pi = \phi'(z)\, dz = \theta\, \phi'(\theta q)\, dq$, so payments integrate to $\int_0^q \theta\, \phi'(\theta s)\, ds = \phi(\theta q)$ for a sole supplier—the case Proposition 1 delivers on strict frontiers; with interleaved suppliers aggregate payments still telescope but individual attribution is path dependent. A participant's lifetime PnL over lifetime tokens is then an average of true marginal products—θ_m times the deployment-weighted average of $\phi'(z)$ it encountered. Whether that average ranks models depends on who chose the deployment, at what scarcity, and on the shape of ϕ. At a common shadow price, the average is an inverse elasticity.

Lemma 2 (PnL per token is an inverse elasticity)**.** *Let $\varepsilon_\phi(z) = z\, \phi'(z)/\phi(z)$ denote the elasticity of task value with respect to effective compute. On any active task with productivity θ at shadow price λ,*

$$\frac{AP(\theta; \lambda)}{\lambda} = \frac{1}{\varepsilon_\phi(z)}, \qquad \theta\, \phi'(z) = \lambda, \tag{39}$$

and $z(\theta, \lambda)$ is increasing in θ. Hence AP ranks θ at a common λ if and only if ε_ϕ is decreasing on the relevant range.

Proof. $AP = \phi(z)/q = \theta\, \phi(z)/z$ and $\lambda = \theta\, \phi'(z)$ give $AP/\lambda = \phi(z)/(z\, \phi'(z))$, and $z = (\phi')^{-1}(\lambda/\theta)$ increases in θ since ϕ' is decreasing. □

Corollary 1 (Log production ranks)**.** *For $\phi(z) = \log(1 + z)$, $\varepsilon_\phi(z) = z/\big((1 + z)\log(1 + z)\big)$ is strictly decreasing, since $\log(1 + z) < z$; hence AP is strictly increasing in θ, explicitly $AP = \lambda\, g(\theta/\lambda)$ with $g(r) = r \log r/(r - 1)$ and $g'(r) = (r - 1 - \log r)/(r - 1)^2 > 0$.*

Concavity alone does not deliver the ranking. The power family $\phi(z) = z^\alpha$, $0 < \alpha < 1$, satisfies every assumption yet has constant elasticity, so $AP/\lambda = 1/\alpha$ for every θ: competitive PnL per token then carries no ranking information at all. And $\phi(z) = z^{1/4} + z^{3/4}$ is increasing and strictly concave with increasing elasticity, so it reverses the ranking under fully competitive deployment at a common shadow price. Declining production elasticity is therefore the substantive economic condition under which inframarginal PnL per token reveals productivity, and whether the elasticity of alpha with respect to compute declines with crowding is an empirical question.

Under log production the margin is the same λ for every survivor, but the inframarginal rents are not, and at a common λ they are monotone in θ. The conditioning on λ is essential: a lifetime quotient that mixes dates or markets with different shadow prices can invert under fully competitive deployment ($\theta_A = 100$ earned where $\lambda = 0.01$ averages $0.01\, g(10^4) \approx 0.09$ per token, while $\theta_B = 2$ earned where $\lambda = 1$ averages $2 \log 2 \approx 1.39$). The repair is to deflate: $AP/\lambda = g(\theta/\lambda)$ ranks relative productivity monotonically—raw PnL per token is a nominal return, and the shadow value of compute is its deflator. Alternatively, when compute is elastically supplied at unit dollar cost (the elastic-supply remark of Section 5), $\lambda = 1$ across unconstrained competitive environments and $AP(\theta) = g(\theta)$ is directly comparable; deflation is needed only where compute is capacity constrained or its real price moves. Note also that when a model serves many tasks its quotient is a token-weighted portfolio of $\lambda\, g(\theta_m(B)/\lambda)$: economic fitness is conditional on both model and task, and a single number summarizes the pieces of

the frontier a model occupies rather than an intrinsic scalar quality. Cross-task comparability further presumes a common, or suitably normalized, response function ϕ across tasks: with task-specific ϕ_B, deflated quotients from different tasks reflect different elasticities as well as different productivities. The two subsections that follow describe the two further ways the comparison degrades.

11.2 Realized marginal PnL per token is equalized, hence uninformative

By (8), every surviving model–task pair earns exactly λ. An econometrician observing realized marginal returns among active technologies sees a flat cross-section: the statistic identifies the shadow price of compute, not relative model quality (Remark 2). Differences in θ show up not in survivors' realized returns but in *who survives where and how much compute they absorb*—quantities, not prices.

11.3 Administered deployment can invert the lifetime quotient

Lemma 2 requires the deployment to be the competitive one. When deployment is administered instead—scale chosen by fiat rather than by itinerant compute—the same quotient can reverse rankings even within one contemporaneous environment, because an administered allocation need not satisfy the common-λ first-order conditions. On a single task, suppressing μ, the average and marginal products of compute run through a technology deployed at scale q are

$$AP(q) = \frac{\phi(\theta q)}{q}, \qquad MP(q) = \theta\, \phi'(\theta q), \tag{40}$$

and concavity gives $AP(q) > MP(q)$ for all $q > 0$, with AP falling in q. With $\phi(z) = \log(1+z)$, route a superior technology A with $\theta_A = 2$ onto a task at scale $q_A = 100$ while an inferior technology B with $\theta_B = 1$ sits at $q_B = 1$. Then

$$AP_A = \frac{\log 201}{100} \approx 0.053, \qquad AP_B = \log 2 \approx 0.693, \tag{41}$$

so the quotient ranks the technology with *half* the productivity an order of magnitude higher, purely because A is over-deployed deep into its diminishing-returns region. Over-deployment is not exotic: a laboratory routing all of its own traffic to its own generalist administers exactly this experiment, and buries a superior technology's average under its own crowding.

11.4 What is a valid estimate

The empirical object that does estimate the theory's selection variable is the *incremental* return

$$\frac{\Delta \mathrm{PnL}}{\Delta T}, \tag{42}$$

measured under the following disciplines, each mandated by a piece of the model:

1. *Small increments.* ΔT small relative to deployed scale, so that the ratio approximates MP, not AP.
2. *Matched task.* Both models evaluated on the same task cell B (task-conditionality of $\theta_m(B)$).

3. *Matched congestion.* Evaluations at a common effective-effort state z_B, since the comparison object (10) holds z fixed; comparing a model attacking a crowded opportunity with one attacking a fresh opportunity measures crowding, not technology.

4. *Standardized compute and PnL.* Tokens denominated in compute cost, and PnL generated under a fixed capital, risk, and execution protocol (Remark 1).

Under these conditions $\Delta\text{PnL}/\Delta T \approx \theta_m(B)\,\phi'(z_B)$, and comparisons across models at matched (B, z_B) recover the ranking of $\theta_m(B)$—the ranking that determines execution. Under competitive deployment the quotient needs only shadow-price conditioning (Lemma 2); the four disciplines above are what restore validity when it is not.

11.5 The intensive and the extensive margin

PnL per token is an intensive-margin statistic: it measures value per token on the niches a model occupies, conditional on operating there. Production share also requires the extensive margin, the economic mass of those niches: Theorem 1 runs through $\mu(E_t)$, and a model with $\theta = 100$ on a niche of measure 10^{-9} combines spectacular PnL per token with negligible share. Commercial survival requires more still. Under log production an owned task yields inframarginal rent

$$R(\theta, \lambda) = \phi(\theta q) - \lambda q = \log \frac{\theta}{\lambda} - 1 + \frac{\lambda}{\theta} \tag{43}$$

above the shadow opportunity cost of its compute; with compute elastically supplied at unit cash cost (the elastic-supply remark of Section 5), $\lambda = 1$ and $R(\theta) = \log\theta - 1 + 1/\theta$ is operating rent above actual compute cost. It is total rent integrated over owned tasks, not average PnL per token, that must eventually cover a technology's fixed development and maintenance cost. The chain runs: task productivity, PnL per token, allocation, total rents, commercial survival. This paper establishes the middle of the chain; the final arrow is the entry model of Remark 5.

11.6 The central claim, precisely stated

Raw capability has none of the relevant equilibrium properties. A model can have $b_G(x) > b_S(x)$ everywhere and nevertheless lie strictly below the productivity frontier everywhere, as in Section 7; with sufficiently fine subdivision, the frontier comparison determines who executes the work.

The claim this paper establishes is:

Survival is determined by task-conditional marginal economic productivity per standardized unit of compute. A competitive prediction market pays exactly that margin, so average PnL per token equals the shadow price of compute times the inverse elasticity of task value with respect to effective compute, and at a common shadow price ranks task-level productivity when that elasticity declines with crowding; the statistic degrades insofar as deployment is administered or scarcity regimes are mixed.

This statement says when a PnL-per-token league table is a valid fitness measure—competitive deployment—and what contaminates it: administered scale, unmatched task mix and congestion, non-standardized compute. For technological survival, the economically fundamental question is not *how capable is the model?* but *on what part of an increasingly subdivided production chain does this model remain the highest-value use of compute?*